

Introduction to the Quantitative Analysis of Textual Data Using quanteda *

Kenneth Benoit and Paul Nulty

June 3, 2014

1 Introduction: The Rationale for quanteda

quanteda is an R package designed to simplify the process of quantitative analysis of text from start to finish, making it possible to turn texts into a structured corpus, convert this corpus into a quantitative matrix of features extracted from the texts, and to perform a variety of quantitative analyses on this matrix. The object is inference about the data contained in the texts, whether this means describing characteristics of the texts, inferring quantities of interests about the texts of their authors, or determining the tone or topics contained in the texts. The emphasis of quanteda is on *simplicity*: creating a corpus to manage texts and variables attached to these texts in a straightforward way, and providing powerful tools to extract features from this corpus that can be analyzed using quantitative techniques.

The tools for getting texts into a corpus object include:

- loading texts from directories of individual files
- loading texts “manually” by inserting them into a corpus using helper functions
- managing text encodings and conversions from source files into corpus texts
- attaching variables to each text that can be used for grouping, reorganizing a corpus, or simply recording additional information to supplement quantitative analyses with non-textual data
- recording meta-data about the sources and creation details for the corpus.

The tools for working with a corpus include:

- summarizing the corpus in terms of its language units
- reshaping the corpus into smaller units or more aggregated units
- adding to or extracting subsets of a corpus
- resampling texts of the corpus, for example for use in non-parametric bootstrapping of the texts (for an example, see ?)

*This research was supported by the European Research Council grant ERC-2011-StG 283794-QUANTESS. Code contributors to the project include Alex Herzog, William Lowe, and Kohei Watanabe.

- Easy extraction and saving, as a new data frame or corpus, key words in context (KWIC)

For extracting features from a corpus, *quanteda* provides the following tools:

- extraction of word types
- extraction of word n -grams
- extraction of dictionary entries from user-defined dictionaries
- feature selection through
 - stemming
 - random selection
 - document frequency
 - word frequency
 - and a variety of options for cleaning word types, such as capitalization and rules for handling punctuation.

For analyzing the resulting *document-feature* matrix created when features are abstracted from a corpus, *quanteda* provides:

- scaling models, such as the Poisson scaling model or Wordscores
- nonparametric visualization, such as correspondence analysis
- topic models, such as LDA
- classifiers, such as Naive Bayes or k -nearest neighbour
- sentiment analysis, using dictionaries

quanteda is hardly unique in providing facilities for working with text – the excellent *tm* package already provides many of the features we have described. *quanteda* is designed to complement those packages, as well to simplify the implementation of the text-to-analysis workflow. *quanteda* corpus structures are simpler objects than in *tm*, as are the document-feature matrix objects from *quanteda*, compared to the sparse matrix implementation found in *tm*. However, there is no need to choose only one package, since we provide translator functions from one matrix or corpus object to the other in *quanteda*.

This vignette is designed to introduce you to *quanteda* as well as provide a tutorial overview of its features.

2 Installing *quanteda*

The code for the *quanteda* package currently resides on <http://github/kbenoit/quanteda>. From an Internet-connected computer, you can install the package directly using the *devtools* package:

```
library(devtools)
if (!require(quanteda)) install_github("quanteda", username = "kbenoit")

## Loading required package: quanteda
```

For other branches, for instance if you wish to install the dev branch (containing work in progress) rather than the master, you should instead run

```
install_github("quanteda", username = "kbenoit", ref = "dev")
```

3 Creating a corpus

3.1 Loading Documents into Quanteda

From a directory of files

A very common source of files for creating a corpus will be a set of text files found on a local (or remote) directory. To load in a set of these files, we will load a corpus from a set of text files using information on attributes of the text that have been conveniently stored in the text document's filename (separated by under-scores). For example, for our corpus of Irish budget speeches, the filename `2010_BUDGET_03_Joan_Burton_LAB.txt` tells us the year of the speech (2010), the type ("BUDGET"), a serial number (03), the first and last name of the speaker, and a party label ("LAB" for Labour).

To load this into a corpus object, we will use the `corpusFromFilenames` function, supplying a vector of attribute labels that correspond with the elements of the filename.

```
library(quanteda)
textfile <- "https://github.com/kbenoit/quanteda/blob/dev/texts/irishbudgets2010.zip?raw=true"
download.file(textfile, basename(textfile), method = "curl", extra = "-L") # download this file
unzip(basename(textfile)) # unzip the file
attNames <- c("year", "debate", "number", "firstname", "surname", "party")
ieBudgets2010 <- corpusFromFilenames("budget_2010", c("year", "debate", "no", "fname",
  "speaker", "party"), sep = "_")
```

This creates a new `quanteda` corpus object where each text has been associated values for its attribute types extracted from the filename:

```
summary(ieBudgets2010)

## Corpus object contains 14 texts.
##
##               Texts Types Tokens Sentences year debate
## 2010_BUDGET_01_Brian_Lenihan_FF.txt    1655    7799     390 2010 BUDGET
## 2010_BUDGET_02_Richard_Bruton_FG.txt     956    4058     222 2010 BUDGET
## 2010_BUDGET_03_Joan_Burton_LAB.txt    1485    5770     329 2010 BUDGET
## 2010_BUDGET_04_Arthur_Morgan_SF.txt    1463    6481     349 2010 BUDGET
## 2010_BUDGET_05_Brian_Cowen_FF.txt    1473    5880     262 2010 BUDGET
## 2010_BUDGET_06_Enda_Kenny_FG.txt    1066    3875     161 2010 BUDGET
## 2010_BUDGET_07_Kieran_ODonnell_FG.txt   614    2066     141 2010 BUDGET
## 2010_BUDGET_08_Eamon_Gilmore_LAB.txt  1098    3800     208 2010 BUDGET
## 2010_BUDGET_09_Michael_Higgins_LAB.txt  447    1136      49 2010 BUDGET
```

```

##      2010_BUDGET_10_Ruairi_Quinn_LAB.txt    418    1177        60 2010 BUDGET
##      2010_BUDGET_11_John_Gormley_Green.txt  363     929        49 2010 BUDGET
##      2010_BUDGET_12_Eamon_Ryan_Green.txt   482    1513        90 2010 BUDGET
##      2010_BUDGET_13_Ciaran_Cuffe_Green.txt  423    1143        48 2010 BUDGET
## 2010_BUDGET_14_Caoimhghin_OCaolain_SF.txt 1055    3654       194 2010 BUDGET
## no      fname  speaker party
## 14 Caoimhghin OCaolain  SF
## 13 Ciaran      Cuffe  Green
## 12 Eamon       Ryan  Green
## 11 John        Gormley Green
## 10 Ruairi      Quinn  LAB
## 09 Michael     Higgins LAB
## 08 Eamon       Gilmore LAB
## 07 Kieran      ODonnell FG
## 06 Enda        Kenny  FG
## 05 Brian       Cowen  FF
## 04 Arthur      Morgan  SF
## 03 Joan        Burton  LAB
## 02 Richard     Bruton  FG
## 01 Brian       Lenihan FF
##
## Source: /home/paul/Dropbox/code/quanteda/vignettes/* on x86_64 by paul.
## Created: Tue Jun  3 18:26:20 2014.
## Notes:  NA.

```

From a vector of texts

3.2 Adding Information to a corpus

Adding new texts

Adding new text attributes

3.3 Translating a quanteda corpus into other formats

Importing from QDAMiner

Importing to and exporting from tm

4 Manipulating a corpus

5 Extracting Features

6 Analyzing a document-feature matrix

References